Generic noun phrases in child speech

Samarth Mehrotra (samarth.1397@gmail.com)
Department of Computer Science, Birla Institute of Technology and Science
Zuarinagar, Goa 403726 India

Amy Perfors (amy.perfors@unimelb.edu.au)
Department of Psychological Sciences, University of Melbourne
Redmond Barry Building, VIC 3010 Australia

Abstract

A wealth of developmental evidence suggests that children essentialise natural kind but not artifact categories, and that both adults and children use generic language less with artifacts as well (Gelman, 2003). Here we further explore the latter result using a novel model for generic identification. We apply our model to a much larger dataset than before, consisting of 26 CHILDES corpora of naturalistic speech involving children at a variety of ages and in a variety of contexts. We found no consistent preference for generic usage in animates over artifacts. Follow-up analyses indicate that this result was probably driven by our inclusion of a wider variety of nouns into our dataset than previous work.
Keywords: essentialism; generics; development; language

Introduction

Psychological essentialism refers to the intuitive belief that many categories have a hidden essence which gives the objects in those categories their identity. Essentialised categories have sharp boundaries, are discovered rather than invented, and have properties that are inherent in some way (e.g., Gelman, 2003). From an early age children behave in ways that are consistent with having essentialist beliefs. This is evident in how they use category information to support induction (Gelman & Markman, 1986) and make predictions about innate potential (Gelman & Wellman, 1991) and identity in the face of transformation (Keil, 1989), among others.

Although there is robust evidence that people essentialise natural kinds, we do not appear to essentialise artifact categories (e.g., Sloman & Malt, 2003). Artifacts do not retain their identity even when transformed (Keil, 1989), often have fuzzy category boundaries (Estes, 2003), and have different insides than animals do (Simons & Keil, 1995).

To what extent is this difference between artifact and natural kinds learned from or supported by environmental differences? One way to answer this question is by investigating one possible source of environmental influence: the use of generic noun phrases (e.g., *Owls sleep during the day* or *Books are heavy*). Generics communicate properties about categories as a whole rather than individuals, and both adults and children appear to make more essentialised inferences when generics are used (Rhodes, Leslie, & Tworek, 2012). Moreover, in a variety of experimental contexts, both children and adults produce generics more often for animals than for artifacts (Gelman & Tardif, 1998; Gelman, Coley, Rosengren, Hartman, & Pappas, 1998; Goldin-Meadow, Gelman, & Mylander, 2005; Brandone & Gelman, 2013). This is highly suggestive that environmental input in the form of generic language usage may play a role in children's early acquisition of essentialised beliefs.

However, the generality of these studies are limited somewhat because they all involved highly structured tasks, often with stimuli specifically created for the experiment. To our knowledge only one study has explored truly *naturalistic* generic language use. Gelman, Sarnecka, and Flukes (2008) hand-coded six corpora for generic language use and found the same bias toward generics in animates over artifacts.

Our work here builds on and extends this research by presenting an automatic model of generic identification. After validating its performance against several external metrics, we apply it to 26 different CHILDES corpora (including the six original ones). Our goal with this larger dataset was to learn more about the range of variation in generic usage in natural speech with children. Are generics used less with artifacts for all corpora, at all ages, and for all speakers? Do the patterns in generic usage support the possibility that psychological essentialism may reflect (or lead to) the statistics of generic speech in the linguistic environment?

Method

The first contribution of our work is the creation of a novel model that can automatically identify generic noun phrases based only on syntactic information. We describe it here.

Model

Although several models for the automatic identification of generic noun phrases exist, they are not ideal for our purposes. For instance, Reiter and Frank (2010) use a Bayesian Network model that relies on a feature set consisting of a large range of both the syntactic and semantic features of the noun itself as well as the clause it is contained in. Example syntactic features include COUNTABILITY, NUMBER, and PART OF SPEECH, while semantic features include SENSE and GRANULARITY. Friedrich and Pinkal (2015) use a conditional random field to label sequences but rely on a similar range of features, both syntactic and semantic.

The reliance on semantic as well as syntactic features is not a problem in general, but does pose an issue for us since our central questions focus on the semantic properties of generic nouns. Do they tend to be animates, artefacts, or something else? We cannot answer this question with a model that identifies generics using semantic features, since any results might emerge due to biases in how the model uses that semantic in-

Word	Part of speech	Dependency label
Elephants	noun	nsubj
do	verb	aux
not	adv	neg
eat	verb	ROOT
birds	noun	dobj
.	punct	punct

Table 1: Example sentence along with the two features used by our model: part of speech and dependency label, which indicates the role each word plays in the syntactic structure.

formation rather than actual distributional properties of the language. We therefore developed a new model of our own which relies only on syntactic features.

Structure Our model is a deep neural network classifier which makes decisions about noun phrases based on their syntactic properties as well as the syntactic properties of other words in the same clause. It therefore incorporates a notion of (local) context: an important consideration when identifying generics because the same word may or may not be a generic depending on how it is used. For instance, the word "dogs" in the sentence *Dogs like to bark* is generic, but the same word in the sentence *Dogs at Pat's house like to bark* is not.

Our classifier was constructed by stacking two different kinds of neural network units together. The first, Long Short-Term Memory (LSTM) units, are especially appropriate to classifying sequence-based data such as words in a sentence, and are widely used in many natural language applications (Hochreiter & Schmidhuber, 1997). We also used Gated Recurrent Units (GRUs) which are similar to LSTMs but often achieve higher performance on smaller datasets like ours. Our model consisted of seven different independently-trained architectures which varied from each other in the dimensionality of the units as well as in how they were stacked.[1]

All of the architectures had a final, fully-connected layer with a softmax activation function which performed the classification task. Each architecture yielded one decision for each noun (generic vs not-generic) and model decisions were made by taking the majority vote among the seven.

Input Our model required two kinds of syntactic information for each of the words in our corpora: the part of speech as well as the dependency label it was associated with in the dependency parse tree. Table 1 illustrates these features for an example sentence. In order to extract this information, we used a number of standard state-of-the art natural language processing tools. We first segmented each of the nouns and their corresponding clauses out of each sentence using the discourse parser SPADE (Soricut & Marcu, 2003). Each word was then assigned a dependency label using the Stanford Dependency Parser (Chen & Manning, 2014) and then tagged with the appropriate part of speech (Toutanova, Klein, Manning, & Singer, 2003).

[1]Our anonymised supplementary materials describe the structure of the architectures: `https://tinyurl.com/ybwg88h5` .

Model	Accuracy	F-score
Reiter and Frank (2010)	71.7	72.3
Friedrich and Pinkal (2015)	79.1	78.8
Our model	76.4	79.3

Table 2: Cross-validation performance on the WikiGenerics dataset. Our model achieves similar performance to the state-of-the-art. Accuracy reflects the total percentage of correct predictions (generics classified as generics, and non-generics as non-generics) while F-score is the harmonic mean of precision and recall, as calculated in Friedrich and Pinkal (2015).

Pronouns posed an interesting dilemma, because they make up a reasonable proportion of all nouns yet cannot be accurately classified for their genericity without determining their referent. For instance, the word "they" in the sentence *Watch out for the piranhas in that fish tank; they bite* is not generic, whereas the word "they" in *I hate mosquitoes; they bite* is generic. We addressed this issue by resolving the coreference of each pronoun using a standard coreference resolution system (Clark & Manning, 2016), and then assigning the genericity of the pronoun to be the same as its referent.

Using these part of speech and dependency features, we created input vectors for our model that corresponded to each noun along with the sequence of words in the clause. This means that for each noun, the model was given not just the noun but also all of the words in the NP it was part of and all of the words in the clause that contained that NP. Each input vector was a concatenation of two vectors consisting of the part-of-speech tag and the dependency label. The model thus used all of the words in the sequence to make a decision about each noun, not just the words that came before it.

Training and validation Each of our seven architectures was trained independently using a weighted categorical cross entropy loss function, which we optimised using the Adam optimiser (Kingma & Ba, 2014). Our loss function weighted the error associated with classifying a non-generic statement as generic (false positive) 1.5 times more than the error associated with classifying a generic statement as non-generic (false negative). By using such a weighted error function, we ensured that the classifier was conservative in its classification of generics, marking a noun as a generic only when it was very confident. This helped to ensure that our model was not overestimating the proportion of generic words.

Before applying our model to CHILDES corpora, we validated its performance in two ways. First we calculated its accuracy and F-score on the WikiGenerics dataset created by Friedrich and Pinkal (2015). This dataset consists of examples from 102 documents from Wikipedia covering a wide variety of topics including animals, games, medicine, music, politics, science, and people, among others. The texts were hand-annotated for genericity by three computational linguists, with contested annotations decided by majority vote. We tested our model using as leave-one-out cross validation strategy. In each cross validation step, examples from 101 of the 102 texts were used for training and the model was tested on the remaining one. The results, shown in Table 2, show

that despite relying on a much smaller range of features our model performed as well as the two best-performing models of generic identification.[2]

Although this level of performance is reassuring, it is not necessarily the case that high performance on a dataset consisting of Wikipedia articles means that the model can accurately identify generics in corpora of child with children. As a second validation of model performance, we thus tested its accuracy against the genericity judgments reported in Gelman et al. (2008).[3] The data we had access to consisted of all of the nouns (in the child speech only) in their six corpora that they coded as generic. Our model had a 88% true positive rate on this data: 88% of the items that they coded as generic were coded as generic by our model. We do not have the list of nouns that they coded as non-generic, but on the assumption that any nouns not coded as generic would have been coded as non-generic, this gives our model an accuracy of 96.8 and an F-score of 81.2 against their gold standard.

CHILDES Datasets We applied our model to 26 different corpora from the CHILDES database (MacWhinney, 2000). The corpora, which are listed in full in the supplemental materials, include the six corpora from Gelman et al. (2008) as well as twenty additional corpora made up of natural conversations between children and adults in English (American or UK). All corpora include both adult and child speech except one (Sawyer) which contained only child speech. Because we were interested in the statistics of language in naturalistic situations, we excluded studies in which children were given a structured task or played with a restricted set of toys.

The supplemental materials list all corpora in detail, but in general, the children ranged in age from less than one year to over five years of age. Given the difficulty in identifying generic usage when grammatical abilities are limited, we excluded all child speech from children less than two years old. However, we do include adult speech to these children because one of our goals with this work is to better understand the distributional properties of the linguistic input they receive at all ages. Our full corpus of child speech contained 1,057,807 utterances total and the corpus of adult speech contained 1,595,305 utterances.

Results

Our first question is about the prevalence of generic speech as a function of age. For the child corpora, we can ask when children begin producing generics. For the adult corpora, we can ask whether adult speech is rich in generics from an early age, and whether there are developmental trends in generic usage. We thus calculated the proportion of generic utterances at different age ranges, coding an utterance as generic if any noun in it was classified as generic. Our results are shown in Figure 1, plotted alongside similar data from Gelman et al. (2008). Although we show more generic usage overall than did Gelman et al. (2008), the patterns are remarkably similar. Children's production increases rapidly over the early years of development, with them producing generics as soon as they have the grammatical capacity. In the early years, adult production is consistently higher than children's, but it then levels off at later ages until they converge. We consider reasons that we estimate more generics in the Discussion.

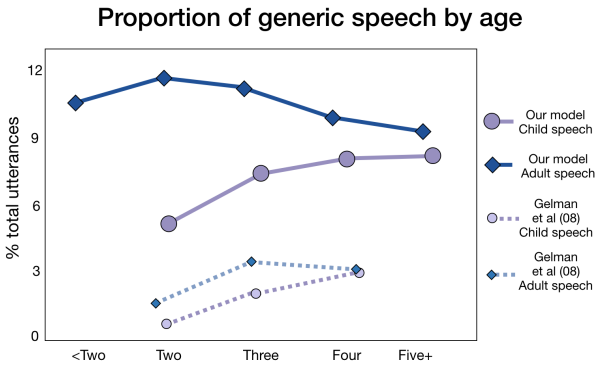

Figure 1: Proportion of generic speech by age. The overall percentage of all utterances coded as generic in our corpora (solid line), broken down by child and adult speech (purple and blue, respectively). For comparison, we plot analogous results from Gelman et al. (2008) with the dotted line. Although we estimated more total generics than they did, the qualitative patterns over development and between child and adult speech are extremely similar.

The primary question motivating this work was how generic usage differs between different kinds of nouns. Do animates, which both children and adults essentialise more, occur more often in generic speech than artefacts, which are essentialised less? In order to investigate this question we had to assign each of the nouns in our corpus to the appropriate category. We accomplished this based on the categories in WordNet, a widely-used lexical database for English. WordNet contains 22 different noun categories, including animals, artifacts, and people as well as feelings, communications, plants, motives, substances, time, and more.

We classified all of our nouns into the four categories used by Gelman et al. (2008): animates, artifacts, food, and other. The artifact and food categories correspond straightforwardly to equivalent categories in WordNet. We constructed our animates category by combining the WordNet animal and person categories, and classified everything else as other. If a word was associated with multiple WordNet categories, we used the Lesk Algorithm to determine which one to assign it to. This algorithm uses the words in the surrounding context to determine the appropriate classification. For instance, the word *fish* would be classified as an animal if it was surrounded by words like *swim* or *water* and as a food if it was surrounded by words like *eat* or *cook*.

What kinds of noun categories do people talk about more, and does this distribution vary between adults and children or by whether generics or non-generics are involved? To answer this question, Figure 2 plots the percentage of each of the four noun categories within generics and non-generics, re-

[2]These numbers are as reported in Friedrich and Pinkal (2015). We did not re-implement their models.

[3]We would like to thank Susan Gelman, who graciously provided this data upon request.

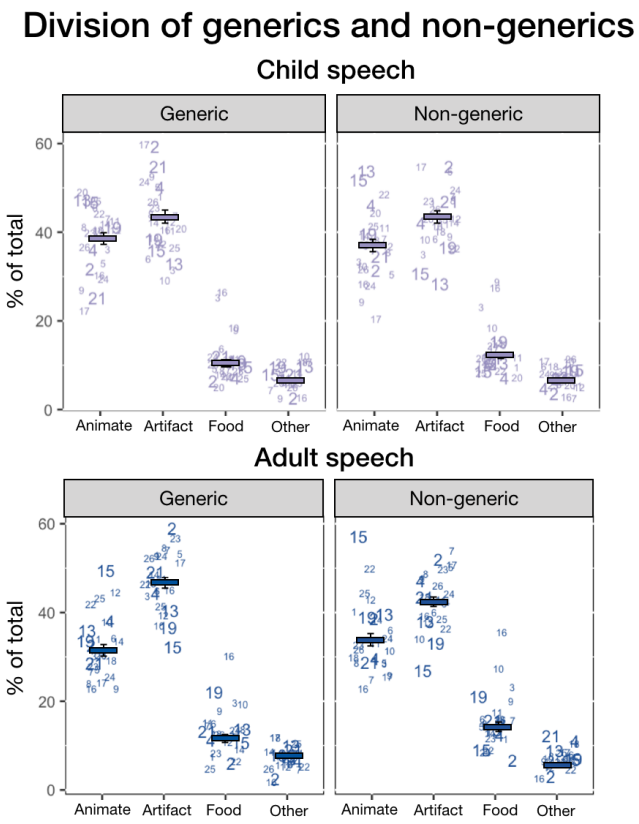

Figure 2: Proportion of generic and non-generic speech across categories. This figure shows the distribution of speech across the four noun categories, for both children (purple) and adults (blue). Lines show the mean when averaged by corpus; error bars indicate standard error. The numbers correspond to the relevant measure for each of the 26 corpora. The corpora from Gelman et al. (2008) are slightly larger and correspond to numbers 2, 4, 13, 15, 19, and 21. It is evident that there is high variability between corpora, but for the most part both children and adults speak about artifacts more often and that there is little difference between generics and non-generics in how they are distributed amongst the four noun categories.

spectively. The left panel thus shows the percentage of all of the generic nouns that are animates, artifacts, foods, or other; the right panel shows the same breakdown out of all of the non-generic nouns. We illustrate the variability in this distribution by plotting the results for each of the corpora individually. It is evident that there is substantial variability overall, and that at least some of that variability is corpus-specific: the correlation between adult and child speech by corpus is $r = 0.95$. This probably largely reflects the fact that children and adults co-create one another's linguistic environment.

This analysis also demonstrates that in general both children and adults talk about artifacts slightly more often than animates.[4] There is also no difference in the distribution of

[4]Bayesian t-test comparing artifact to animate percentage: For child generics, $BF_{10} = 2.4$ weakly in favour of a model that includes noun type; for child non-generics, $BF_{10} = 7.7$ moderately in favour. For adult generics, $BF_{10} > 10^6$ in favour of a model that includes nountype; for adult non-generics, $BF_{10} = 111$ in favour. All Bayesian analyses used the BayesFactor package in R (version 3.4.4) and compared the model of interest to an intercept-only null model. In also cases we also ran analogous frequentist tests, which always returned qualitatively similar results.

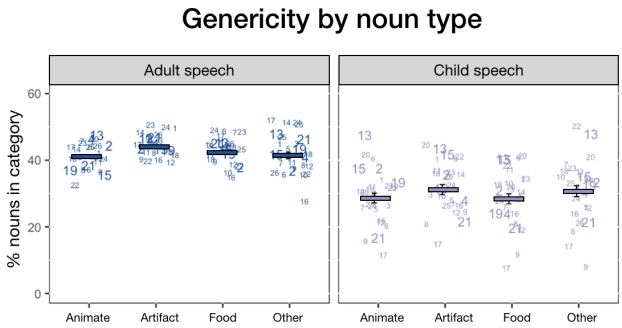

Figure 3: Proportion of generic speech within each noun category. For each of the four categories, this figure shows how often nouns in that category were generic. The large transparent bars indicate the aggregate proportion over all corpora, while the small boxes with error bars show the mean when averaged by corpus. The numbers correspond to the relevant measure for each of the 26 corpora. The corpora from Gelman et al. (2008) are slightly larger and correspond to numbers 2, 4, 13, 15, 19, and 21. There is high variability between corpora (especially for children). However, there is little difference in the pattern of generic usage across noun categories.

speech across noun categories as a function of genericity or speaker.[5] Generics and non-generics have similar distributions across different kinds of nouns, and this holds regardless of whether the speakers are adults or children.

Another way to explore the issue of whether children or adults use generics differently for different categories is to condition on category rather than on genericity. Figure 3 thus shows, for each of the four noun categories, what proportion of time it occurs as a generic in both child and adult speech. Although children are much more variable, we still see little difference in generic usage between noun categories. However, adults were more likely to use generics for artifacts than animates, as well as more overall.[6]

These results are rather surprising, since previous work has suggested that generics tend to be used more often with animate categories. What is going on?

One possibility might be that the six corpora used by Gelman et al. (2008) were outliers in some way relative to our larger set of 26. In order to investigate this possibility, we calculate how many corpora used a higher percentage of animate nouns than artifact nouns as generics. On this measure, the corpora from Gelman et al. (2008) appear to be slight outliers relative to the others. Of the 25 corpora with adult speech, only six used generics more with animates and three of those six were theirs: Bloom (2), Brown (4), and Kuczaj (13). Of the 26 with child speech, six used generics more with animates and four were theirs: 2, 4, 13, and Sachs (19).

[5]Bayesian ANOVA: $BF_{01} = 10$ for the null model over a model including genericity and $BF_{01} = 10$ for the null over a model including speaker. This indicates strong support for the null model.

[6]Bayesian ANOVA: $BF_{01} = 14.3$ favouring the null model over a model including noun category; $BF_{10} > 10^6$ favoring a model including speaker. Bayesian t-test comparing artifact to animate generic percentage: for child speech, $BF_{01} = 1.9$ favouring the null model; for adult, $BF_{10} = 6.9$ favouring a model including nountype.

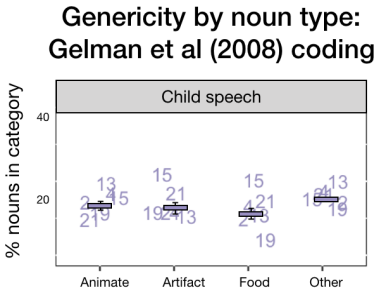

Figure 4: Proportion of generic speech within each noun category using the genericity identifications from Gelman et al. (2008). For each of the four categories, this figure shows how often nouns in that category were generic, using the six corpora and their classifications rather than the classifications from our model. Despite using their classifications, we replicate our previous result, suggesting that the difference between our findings and theirs did not arise due to poor classification performance by our model.

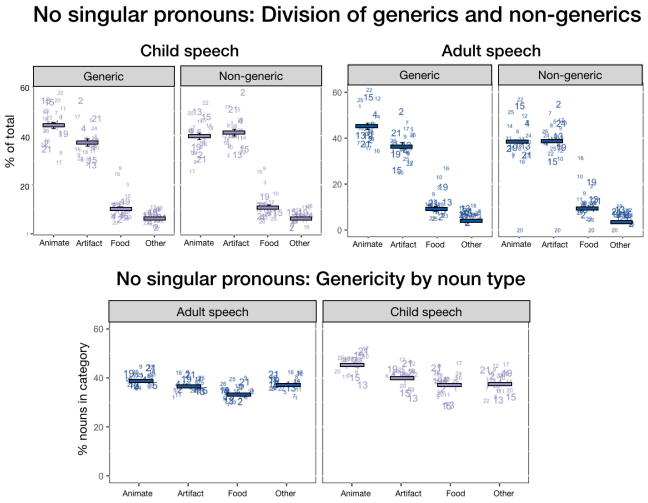

Figure 5: Proportion of generic speech across and within noun categories, on corpora without any singular pronouns. Since Gelman et al. (2008) excluded singular pronouns, we reran our analyses (using our classifications) on our corpora after excluding all singular pronouns. Results are now much more similar to their findings than ours. Generics but not non-generics are used more for animate than artifact categories (top); and for both adults and children, the proportion of generic utterances in animates is higher than in artifacts. This suggests that their exclusion of singular pronouns from the dataset may have driven their results.

These considerations suggest that at least part of the reason our results diverge so markedly from Gelman et al. (2008) is that their corpora were different. However, this cannot be the entire story: the magnitudes of the differences they found are much larger than the magnitudes we found on the same corpora in very similar analyses.

An obvious possibility is that our model is simply classifying many items very differently than they did. Our high accuracy and F-score against their coding scheme suggests that this is not the case, but we were able to test this hypothesis in a much more stringent way as well. For the six corpora in Gelman et al. (2008) that we have their classifications for (child speech only), we took the set of nouns that they identified as generic, assumed that they coded all of the others as non-generic, and applied the same analysis as in Figure 3 to that data. If the difference between our work is because our classifier is coding or identifying items differently than they did, we should find that using their classifications on their corpora replicates their results. However, Figure 4 reveals that we instead replicate our result: there is no difference in generic usage across the four noun categories.[7]

This outcome suggests that the point of divergence between our work and Gelman et al. (2008) must be less due to different decisions about what to code as generic, and more due to different decisions about what nouns to include in the first place. Our analysis included all nouns of any kind, which was straightforward to do since the model could identify them automatically. However, lacking this technology, Gelman et al. (2008) had to process the corpora by hand. They accomplished this by manually identifying potential generics by searching for any bare plurals, plural pronouns, mass nouns, and indefinite singular nouns and then hand-coding that set of nouns as generic (or not). This was justified on the grounds that the vast majority of generics fall into these categories, which is sensible if the goal is to understand the distribution of generics alone. However, if the goal is also to compare to non-generics, it is important to include even those nouns that tend to be non-generic. Their dataset excluded singular pronouns like *he, she, you, I*, and *it*. If singular pronouns tend to "cluster" (for instance, are more likely to be animate and non-generic) then excluding them might result in a mis-estimation of the overall distribution of generics relative to non-generics in different ways for different noun categories.

To test whether the inclusion or exclusion of singular pronouns drove the difference between our results and those of Gelman et al. (2008), we re-ran our original analyses after excluding all singular pronouns from our dataset. As shown in Figure 5, the results now replicate their findings rather than ours. The top panel shows that generics but not non-generics are used more for animate than artifact categories,[8] and the bottom panel shows that for both adults and children, the proportion of generic utterances is higher within animate categories than artifacts.[9] This suggests that Gelman et al. (2008) may have found that animate categories had more generics because they did not count a large number of non-generic animates like *he, she, you,* and *I*. Our other analysis show that once all nouns are included, the proportion of generics across noun categories evens up and if anything favours artifacts.

[7]Bayesian ANOVA: $BF_{01} = 2.6$ for the null over a model including noun category. Bayesian t-test comparing animates to artifacts: $BF_{01} = 2.1$ for the null over a model including noun category.

[8]Bayesian t-test comparing artifact to animate percentage: For child generics, $BF_{10} = 91$ in favour of a model that includes nountype; for child non-generics, $BF_{01} = 2.9$ for the null model. For adult generics, $BF_{10} = 206$ in favour of a model that includes nountype; for adult non-generics, $BF_{01} = 3.5$ for the null model.

[9]Bayesian ANOVA: $BF_{10} > 10^6$ favouring a model including noun category; $BF_{10} = 522888$ favoring a model including speaker. Bayesian t-test comparing artifact to animate generic percentage: for child speech, $BF_{10} = 2529$ favouring a model including nountype; for adult, $BF_{10} = 35$ favouring a model including nountype.

Discussion

This work makes several contributions. First, we present the first fully automatic model for generic identification which uses only synactic features, and demonstrate that it performs well relative to both the state-of-the-art and manual classifications from Gelman et al. (2008). Second, we apply this model to a much larger dataset of child speech than had previously been possible to analyse. Although we replicate the previously-observed developmental trend in generic usage, we find that neither adults nor children use generics more in categories that tend to be essentialised (like animates). Follow-up analyses suggest that our results differ from Gelman et al. (2008) not because of poor classification performance by our model, but primarily because we did not exclude singular pronouns from our dataset (as they did).

A natural question at this point is whether it is better to include singular pronouns or not. Any answer must be conditioned on considerations of what is realistically possible. Given the extreme amount of labour involved in hand-coding corpora, one can reasonably argue that the process for identifying nouns used by Gelman et al. (2008) was a necessary simplification. Other analyses excluded pronouns for other good reasons. For instance, Gelman and Tardif (1998) and Goldin-Meadow et al. (2005) excluded pronouns because of the need to compare English with Mandarin, a pro-drop language. Given these considerations, this too seems reasonable. However, it is possible that this decision is why they as well found a higher proportion of generics for animates.

Overall, we suggest that if the goal is to understand the distribution of generics relative to non-generics in the nouns children hear, it is important to include *all* of the nouns that children hear. Singular pronouns are very common and almost always non-generic; as such, an accurate comparison of generics to non-generics cannot exclude them.

One might also ask why our model identified a larger proportion of generics than previous work did (Figure 1). Part of the reason is probably that a manual identification of generics, as Gelman et al. (2008) had to do, would probably have erred on the side of under-counting them. Another part is that our model appeared to make less conservative choices in some cases. For instance, our model identified many generics that were preceded by the word *the*, as in sentences like *What do bears in the forest do in the day?*. Since our observed developmental trends are very similar and all of our other results hold even when we use the classifications from Gelman et al. (2008), we doubt that our overall higher rate poses a problem.

A final question is what our results mean for our initial question: to what extent does the linguistic environment support the difference in essentialisation of artifact vs animate categories? Our results suggest that this difference is not reflected in differences in generic usage, and thus lends less credence to the possibility that these domin differences in essentialism result from linguistic input. Although this finding is surprising given previous work, one nice aspect of it is that it removes the chicken-and-egg question that otherwise arises: *why* does the linguistic environment have this distribution in the first place? Much remains to be done, but we hope that our model and results offer a useful tool for better understanding how our early biases are shaped by the environment.

References

Brandone, A., & Gelman, S. (2013). Generic language use reveals domain differences in children's expectations about animal and artifact categories. *Cog. Dev.*, *28*, 63–75.

Chen, D., & Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *EMNLP*.

Clark, K., & Manning, C. (2016). Improving coreference resolution by learning entity-level distributed reps. In *ACL*.

Estes, Z. (2003). Domain differences in the structure of artifactual and natural categories. *Mem & Cogn.*, *31*, 199–214.

Friedrich, A., & Pinkal, M. (2015). Discourse-sensitive automatic identification of generic expressions. In *ACL*.

Gelman, S. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford Univ. Press.

Gelman, S., Coley, J., Rosengren, K., Hartman, E., & Pappas, A. (1998). Beyond labeling: The role of maternal input in the acquisition of richly-structured categories. *SRCD*, *63*.

Gelman, S., & Markman, E. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

Gelman, S., Sarnecka, P. G., & Flukes, J. (2008). Generic language in parent-child conversations. *LL&D*, *4*(1), 1–31.

Gelman, S., & Tardif, T. (1998). A cross-linguistic comparison of generic noun phrases in English and Mandarin. *Cognition*, *66*(3), 215–248.

Gelman, S., & Wellman, H. (1991). Insides and essences: early understandings of the nonobvious. *Cogn.*, *38*.

Goldin-Meadow, S., Gelman, S., & Mylander, C. (2005). Expressing generic concepts with and without a language model. *Cognition*, *96*, 109–126.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Keil, F. (1989). *Concepts, kinds, and cogn. devel.* MIT Press.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv:1412.6980*.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum.

Reiter, N., & Frank, A. (2010). Identifying generic noun phrases. In *ACL* (pp. 40–49).

Rhodes, M., Leslie, S., & Tworek, C. (2012). The cultural trans. of social essentialism. *PNAS*, *109*, 13526–13531.

Simons, D., & Keil, F. (1995). An abstract to concrete shift in the development of biological thought: The insides story. *Cognition*, *56*, 129–163.

Sloman, S., & Malt, B. (2003). Artifacts are not ascribed essences, nor are they treated as belonging to kinds. *Language and Cognitive Processes*, *18*, 563–582.

Soricut, R., & Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical info. In *NAACL*.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*.